Kansalliskirjasto
Raportteja ja selvityksiä 2 / 2017

SUKUKIELTEN DIGITOINTIPROJEKTI

Jatkohankkeen loppuraportti
2014–2016

Jussi-Pekka Hakkarainen

Kansalliskirjasto, Helsinki 2017

ISBN 978-951-51-2910-9 (pdf)

KUVAILUSIVU

Julkaisija	Kansalliskirjasto
Julkaisun päivämäärä	24.1.2017
Tekijä(t)	Jussi-Pekka Hakkarainen, projektipäällikkö
Julkaisun nimi	Sukukielten digitointiprojekti Jatkohankkeen loppuraportti, 2014–2016
Julkaisun osat	
Sarjan nimi ja numero	Raportteja ja selvityksiä 2/2017
ISSN **ISBN**	978-951-51-2910-9
URL **URN**	
Kokonaissivumäärä	23 sivua
Kieli	Suomi
Avainsanat – YSA (Suomi)	kielisukulaisuus uralilaiset kielet: http://finto.fi/yso/fi/page/p7464 kielentutkimus: http://finto.fi/yso/fi/page/p21012 kieliteknologia: http://finto.fi/yso/fi/page/p6071 tieteellinen yhteistyö: http://finto.fi/yso/fi/page/p10380 uhanalaiset kielet: http://finto.fi/yso/fi/page/p555 Sukukielten digitointiprojekti Digitization Project of Kindred Languages Проект по оцифровке родственных языков
Tiivistelmä	Raportissa kuvataan Kansalliskirjaston toteuttaman Sukukielten digitointi-projektin toteutustapaa, projektin tuloksia ja vaikutuksia.

Sisältö

TIIVISTELMÄ

Kansalliskirjasto toteutti Koneen Säätiön rahoituksella Sukukielten digitointiprojektin jatkohankkeen vuosina 2014–2016. Sukukielten digitointiprojektin jatkohanke oli osa Koneen Säätiön Kieliohjelmaa, jossa yhtenä keskeisenä tavoitteena oli pienten suomalais-ugrilaisten kielten, suomen sekä Suomen vähemmistökielten dokumentointi ja niiden aseman vahvistaminen. Sukukielten digitointiprojektissa tämä tavoite ymmärrettiin tehtävänä saattaa kieliaineistoja tiedeyhteisön ja muun yhteiskunnan avoimeen käyttöön.

Projektissa tuotettiin uralilaisilla kielillä painettuja aineistoja ja aineiston jalostamisvälineitä kielentutkimuksen ja kansalaistieteen tueksi, Koneen Säätiön Kieliohjelman mukaisesti. Jatkohankkeen aikana digitoitiin ja saatettiin käyttöön lähes 1100 monografia- ja 51 sanomalehti-nimekettä. Monografiasivuja suunnitelman mukaisesti kertyi noin 88 300 ja sanomalehtisivuja noin 72 500. Aineistoja oli yhteensä 14 eri uralilaisella kielellä ja ne on digitoitu Venäjän Kansalliskirjaston kokoelmista.

Digitoitavat aineistot valittiin yhteistyössä kotimaisen tutkijakunnan kanssa ja niiden katsottiin palvelevan niin kotimaista kuin ulkomaista fennougristiikan alan tutkimusta. Projektissa tuotettu aineisto saatettiin sekä tutkijoiden hyödynnettäväksi että avoimeen kansalaiskäyttöön Kansalliskirjaston ylläpitämässä Fenno-Ugrica -kokoelmassa: http://fennougrica.kansalliskirjasto.fi/

Aineistot ovat vapaasti saavutettavissa, sillä Kansalliskirjasto on projektin aikana selvittänyt aineistoihin liittyvät tekijänoikeudelliset kysymykset yhdessä National Library Resourcen ja Venäjän Kansalliskirjaston kanssa. Tämä on mahdollistanut aineistojen avoimen käytön sekä niiden linkittämisen kolmansiin järjestelmiin.

Sukukielten digitointiprojektissa pyrittiin hyödyntämään myös kansalaistieteiden metodeja joukkoistamalla kieliaineistojen oikolukua. Joukkoistamisessa pyrittiin vastavuoroisuuteen, jolloin kansalaistieteilijät ja heidän yhteisönsä hyötyisivät hankkeen tuloksista. Oikolukua voitiin tehdä digitointiprojektissa suunnitellun avoimen lähdekoodin Revizor-editorin avulla.

Sukukielten digitointiprojektin jatkohanke nojasi suurilta osin vuosina 2012–2013 tehdyn pilottihankkeen toimintatapoihin ja käytänteisiin. Ks. Pilottihankkeen loppuraportti: http://www.doria.fi/handle/10024/99181

1 Kielten dokumentointi

Koneen Säätiön myöntämällä apurahalla rahoitettiin vuosina 2014–2016 Sukukielten digitointiprojektin jatkohanke, jonka tavoitteena Kansalliskirjastolla oli kieliaineistojen tuottaminen, dokumentointi sekä talkoistaminen, kielentutkimuksen, -opetuksen ja -elvyttämisen tueksi.

Kielten dokumentointi on ollut viimeisen kahdenkymmen vuoden aikana vilkkaasti debatoitu aihe, kun sen merkitys uhanalaisten kielten säilyttämiselle on ymmärretty. Kielen dokumentointi on siirtymässä pois kielen deskriptiivisestä kuvailusta kohti kokonaisvaltaisempaa ja kieltä säilyttävää dokumentointia. Kansalliskirjastossa ja Sukukielten digitointiprojektissa kielen dokumentaatio ymmärrettiin teknisten ratkaisujen ja aineiston avoimen saatavuuden sekä monikäyttöisyyden kautta. Keskeisiä komponentteja kielen dokumentaatiossa ovat:

1) aineistojen tallentaminen, mukaan lukien siihen liittyvän metadatan tuottaminen
2) kieliaineistojen siirrettävyys
3) lisäarvon tuottaminen mm. annotoimalla, transkriboimalla ja linkittämällä
4) arkistointi ja arkistoidun aineiston avoin saavutettavuus sekä
5) aineiston mobilisointi, eli sen hyödynnettävyys kolmansissa järjestelmissä.

Nämä kielen dokumentoinnin komponentit olivat olennainen osa Sukukielten digitointiprojektissa harjoitettuja käytänteitä. Kielen dokumentoinnin keskeisempiä tavoitteita on, ja myös Sukukielten digitointiprojektissa oli, dokumentaation palauttaminen kieliyhteisön hyödynnettäväksi. Kansalliskirjasto pyrki digitointiprojektien kuluessa luomaan kansalaistieteilijöiden verkoston, jotka ovat toimineet digitoidun aineiston oikolukijoina. Kieliaineistojen korjaamisessa pyrittiin hyödyntämään vuorovaikutuksellisia kansalaistieteiden metodeja. Joukkoistamalla aineistojen oikolukua voidaan yhdistää ammattilaisten osaamista optimoimaan monimutkaistenkin työtehtävien ihmislähtöistä suorittamista kansalaistieteilijöiden voimin. Kansalaistieteilijöiden kanssa tavoitellaan vastavuoroisuutta, jolloin heidän edustamat yhteisöt hyötyvät hankkeen tuloksista.

Kielten dokumentoinnin viisi komponenttia ovat olleet kohtalaisen helposti siirrettävissä Kansalliskirjaston infrastruktuureihin (Fenno-Ugrica -kokoelma). Aineistot on tallennettu verkkokokoelmaan noudattaen bibliografisen kuvailun sääntöjä, mukaan lukien laadukas metadata. Kieliaineistoja on voitu siirtää kolmansiin järjestel-

miin, kiitos aineistoihin kohdistuneen tekijänoikeuksien selvityksen, mikä on mahdollistanut myös aineiston arkistointia ja aineiston avointa saavutettavuutta – kaikki Fenno-Ugricassa julkaistut aineistot ovat sekä kieli-, että tutkijayhteisöjen saavutettavissa. Aineistoille on pyritty tuottamaan lisäarvoa mm. oikolukemalla niitä. Lisäksi aineistoja on voitu hyödyntää useissa kolmansissa järjestelmissä vapaasti.

Yksiselitteisesti voidaankin todeta, että Kansalliskirjaston toiminta, sen infrastruktuuri ja sen tarjoamat palvelut tukevat kielen dokumentointia ja Koneen Säätiön Kieliohjelmassa asetetut tavoitteet täyttyvät.

2 Aineistot

2.1 Aineistojen valinta

Aineistojen valinnassa käytettiin useita kriteereitä, joita Kansalliskirjasto määritteli yhdessä yhteistyökirjastojen ja tutkijakunnan kanssa. Keskeisimpänä valintakriteerinä oli nykykirjakielten synty- ja vakiintumisajankohta. Komin, udmurtin, vuori- ja niittymarin, inkeroisen, ersän, mokšan ja vepsän osalta tämä prosessi tapahtui maailmansotien välisenä aikana, keskeisimmin vuosina 1932–1937. Puhujamäärältään edellä mainittuja kieliä pienempien samojedi- ja obinugrilaisten kirjakielten syntyminen on tapahtunut vasta toisen maailmansodan jälkeen, joskin jo 1930-luvulta voidaan havaita fragmentaarista esim. nenetsillä julkaistua kieliaineistoa.

Sukukielten nykyisille vaalijoille kirjakielen syntykauden aineisto on tärkeää: 1920- ja 1930-luvun uudissanat ja niitä käyttävät tekstit toimivat meidän aikamme kielenkehittäjille yhtä lailla lähdemateriaalina kuin innovaatioiden ja innoituksen lähteenä. Digitoitaviksi esitetyt teokset valittiin siten, että ne yhtäältä kuvastaisivat mahdollisimman hyvin 1920-luvun innovatiivista aikaa mutta toisaalta ilmentävät myös sitä kielipoliittista muutosta, joka tapahtui 1930-luvulla.

Pilottivaiheen kielivalikoimaa (inkeroinen, vepsä, marilaiset ja mordvalaiset kielet) laajennettiin jatkohankkeen aikana. Sukukielten digitointiprojektin jatkohankkeen tuotettiin aineistoja, joita oli julkaistu erityisesti permiläisillä (udmurtti, komi ja komipermjakki), obinugrilaisilla (hanti ja mansi) ja samojedikielillä (nenetsi ja selkuppi). Näillä kielillä painettua aineistoa tuotiin ensi kertaa tuotantoon marilaisten ja mordvalaisten kielten ohessa. Laajempi kielten valikoima tuki osaltaan niin Kieliohjelman puitteissa kuin sen ulkopuolella tehtävää kielentutkimusta, sekä kotimaassa että ulkomailla.

Keskisuurten kielten (komi, udmurtti, ersä, mokša, niitty- ja vuorimari) osalta pyrittiin digitoimaan sekä monografia- että sanomalehtimuotoista aineistoa. Jatkohankkeen kannalta oli tarkoituksenmukaista digitoida sellaista monografia-muotoista kieliaineistoa, jota oli käännetty venäjästä kohdekielelle. Tätä valintaa puolsi ennen kaikkea nk. rinnakkaisnimekkeiden tuottamisen kustannustehokkuus, joita saavutetaan erityisesti aineistoon kohdistuvien tekijänoikeuksien selvittämisessä. Lisäksi tavoitteena oli, että erikielisiä saman teoksen käännöksiä voitaisiin käyttää paralleelitekstien tavoin, jolloin kielten vertailu helpottuisi. Rinnakkaisnimekkeitä, eli venäjänkielisen alkuteoksen käännöksiä, valittiin sellaisilta sanaston alueilta, joita esiintyy vain rajallisesti sanomalehtiteksteissä. Tämän valinnan myötä rinnakkaisnimekkeinä digitoidaan paljon koulu- ja oppikirjoja sekä valistuskirjasia eri opin- ja tieteenaloilta, joita tutkijoiden kanssa määriteltiin yhteensä 27.

Sanomalehtien kohdalla pyrittiin painottamaan ennen kaikkea hyvin vähän digitoituja sekä vaikeasti saavutettavia alueellisia ja periferisiä sanomalehtiä. Valinnan taustalla oli ajatus, että sisällön puolesta keskuksen ulkopuolella sijaitsevien alueiden kieli (paikallislehdet) olisi keskuksen kieltä mielenkiintoisempaa, koska periferiassa voi ilmetä joko keskuksen kielestä poikkeavaa murrevariaatiota tai konservatiivisuutta myös kirjoitetussa kielessä. Paikallisten sanomalehtien digitointia puolsi myös pyrkimys edistää aineiston saavutettavuutta: paikallismateriaaliin keskittymällä tuotiin tutkijoiden ulottuville aikaisemmin huonosti saatavaa materiaalia ja sellaisia sanomalehtiä, joiden asema venäläisten kirjastojen digitointisuunnitelmissa oli joko hyvin marginaalinen tai jopa poissaoleva.

Kieliaineistojen valintoja rajoittavia tekijöitä oli muutamia. Yksi keskeinen rajoitus oli aineiston saavutettavuus Suomesta. Mikäli teos löytyi joko Kansalliskirjaston tai suomalaisten yliopistonkirjastojen kokoelmista, teosta ei otettu mukaan digitoitavaksi, sillä tutkijoiden ajateltiin pääsevän käsiksi aineistoon muutoinkin. Aivan täydelliseen päällekkäisyyksien poistamiseen ei kuitenkaan päästy.

Toinen merkittävä rajaava tekijä on ollut aineistoihin kohdistuvat tekijänoikeudet. Sukukielten digitointiprojektin pilottihankkeen (2012–2013) aikana saatujen kokemusten mukaan ennen vuotta 1941 julkaistua aineistoa voitaisiin saattaa avoimeen tutkija- ja kansalaiskäyttöön verrattain helpostikin, mutta toisen maailmansodan jälkeinen aineisto edellyttäisi väistämättömästi sopimista tekijänoikeudenhaltijoiden kanssa. Aikaisemmat kokemukset tekijänoikeuksien selvittämisen vaikeudesta ohjasi aineistojen valintaa ja siten painotus on vahvasti 1920- ja 1930-luvuilla julkaistuissa aineistoissa.

Sisällölliset näkökohdat aineiston valinnassa puolsivat nimenomaan kirjakielten muotoutumisvaiheisiin liittyvien tekstien sisällyttämistä digitoitavaan materiaaliin. Tämä koski erityisesti keskisuuria kirjakieliä (niittymari, vuorimari, ersä, mokša, udmurtti, komi), joiden tapauksessa murteiden rooli kirjakielen synnyssä on mielenkiintoinen tutkimuskohde. Muutokset kirjakielissä ovat 1950-luvulta lähtien aika vähäisiä verrattuna aikaisempiin vuosikymmeniin, jolloin sota-ajasta sekä sen aiheuttamista ilmiöistä oli jo selvitty ja oli siirrytty vakaan neuvostosysteemin aikaan.

Pienempien ja uhanalaisten kielten osalta tästä vuoden 1941 aikarajasta joustettiin niin usein kuin se oli aineiston merkityksen kannalta perusteltavissa, sillä jatkohankkeessa pyrittiin korkeaan kattavuuteen materiaalin digitoinnissa: lähes kaikki Venäjän Kansalliskirjastossa olevat ennen vuotta 1955 julkaistut monografianimekkeet hantilla, mansilla, inkeroisella, nenetsillä ja selkupilla digitoitiin projektin aikana. Tämä valinta tulee käsittää pyrkimyksenä tukea uhanalaisia kieliä ja kielidiversiteetin edistämisenä.

Ks. liite 1 (Digitoidut aineistot).

2.2 Aineistoihin liittyvät selvitykset, kustannukset ja tekniset ratkaisut

Aineisto digitoitiin lähes kokonaisuudessaan Pietarissa sijaitsevan Venäjän Kansalliskirjaston kokoelmista. Aineistojen julkaiseminen, niistä tuotetun datan ja niiden linkittäminen kolmansiin järjestelmiin ei olisi ollut mahdollista ilman kattavaa tekijänoikeuksien selvitystä. Tekijänoikeuksien selvitys tehtiin Suomen ja Venäjän Kansalliskirjastojen toimeksiannosta Moskovassa. Tekijänoikeuksia selvitti National Library Resource, joka pyrki tavoittamaan tekijänoikeudenhaltijoita Venäjällä.

Jatkohankkeen aikana pyrittiin digitointiyhteistyöhön myös muiden Venäjällä olevien kansalliskirjastojen kanssa. Tavoitteena oli täydentää erityisesti Fenno-Ugrican sanomalehtikokoelmaa paikallisten kirjastojen kausijulkaisukokoelmista, mutta Komin tasavallan kansalliskirjastoa lukuun ottamatta aloitteet eivät konkretisoituneet aidoksi digitointiyhteistyöksi. Syyt siihen, ettei digitointiyhteistyötä saatu käynnistettyä eräiden suomalais-ugrilaisilla alueilla toimivien kansalliskirjastojen kanssa, liittyvät ensisijaisesti vajanaiseen osaamiseen ja sopivan laitteiston puuttumiseen ja toissijaisesti vähäiseen kokemukseen kansainvälisestä yhteistyöstä, mikä osaltaan korotti kynnystä yhteistyön aloittamiseksi.

Aineistojen digitointiin, tekijänoikeuksien selvittämiseen ja Euroopan Unionin ulkopuolisista maista tehtäviin palveluostoihin, joita siis Venäjän Kansalliskirjastossa tehty digitointi ja sanomalehtiaineistojen konservointi oli, budjetoitiin jatkohankkeen alussa 376 000 euroa. Todelliset Venäjältä tehtyjen palveluostojen kustannukset olivat huomattavasti budjetoitua pienemmät, yhteensä 265 000 euroa. Tämä oli seurausta ruplan kurssin rajusta heikkenemisestä hankekauden aikana. Palveluostoista jääneitä varoja käytettiin ennen kaikkea henkilöstön palkkaamiseen. Ks. liite 2 (Apurahahakemus Koneen Säätiölle).

Digitoidut aineistot on saatettu käyttöön Fenno-Ugrica -palvelussa. Fenno-Ugrica on toteutettu DSpace-ohjelmistolla. DSpace on avoimen lähdekoodin sovellus digitaalisten aineistojen hallinnointiin. Kansalliskirjasto käyttää sitä eräiden omien aineistojensa hallintaan sekä tarjoaa sen avulla tuotettuja maksullisia julkaisuarkistopalveluja. Projektin tarpeita varten räätälöidyn DSpace-instanssin avulla edistetään digitoidun aineiston saavutettavuutta ja käyttöä myös ulkomailla, niin kansainvälisen tiedeyhteisön kuin Venäjällä asuvien suomensukuisten kielten puhujien keskuudessa. Fenno-Ugrica toteutettiin myös venäjänkielisenä. Viestinnän ja tunnettuvuuden edistämiseksi Fenno-Ugricalle luotiin räätälöity ja tunnistettava ulkoasu.

Käyttäjille tarjotaan ensisijaisesti PDF-versiot lukukappaleiksi, mutta monografiamuotoisesta aineistoista on valmistettu myös datapaketit, jotka sisältävät kunkin teoksen informaation pitkäaikaissäilytykseen soveltuvissa formaateissa Alto XML, CSV, TXT ja TIFF. Myös loppukäyttäjien, erityisesti tutkijoiden kannalta on ollut

suotavaa, että teosten kuvat on jaettu avoimesti osana kokoelmaa. Tämä on paikannut puutteita, joita Kansalliskirjaston tuottaman data julkaisemisessa ja jakelussa on esiintynyt. Valitettavasti Kansalliskirjastolla ei edelleenkään ole omaa datakatalogiansa, minkä kautta aineisto olisi voitu luotettavasti jakaa, joten Fenno-Ugricaa on käytetty myös tähän tarkoitukseen. Sanomalehtiaineistojen datapaketteja niiden laajuudesta johtuen ei ole jaettu osana Fenno-Ugricaa – sanomalehtiaineistosta on käytettävissä vain PDF-muotoiset digitaaliset käyttökopiot.

Fenno-Ugricassa julkaistu aineisto on linkitetty myös Kansalliskirjaston ylläpitämiin kirjastotietokantoihin, Helkaan ja Melindaan. Aineistot ovat haettavissa Finna-hakupalvelussa. Tavoitteena on ollut parantaa aineiston saavutettavuutta ja edistää digitaalisten aineistojen käyttöä.

2.3 Aineistojen käyttö ja käytön tilastointi

Fenno-Ugrican käyttöä on pyritty seuramaan myös Google Analytics -tilastointipalvelun avulla. Google Analyticsin mukaan Fenno-Ugrica -kokoelmaa on käytetty 108 maasta. Aineistoja on käytetty eniten Venäjältä ja Suomesta. Venäjältä saapuva kokoelman käyttö on lähes 40% (39,91%) kokonaiskäytöstä, kun taas 33,27% kokoelman käytöstä tulee suomalaisista ip-osoitteista.

Kokoelman käyttö on lisääntynyt projektin aikana huomattavasti. Kun vuonna 2013 kokoelmasta ladattiin aineistoja vain 5715 kertaa, niin vuonna 2014 latauksia oli lähes 100 000 kpl. Vuoden 2015 aikana kokoelman käyttö lähes tuplaantui 171 000 lataukseen, kun taas vuoden 2016 aikana kokoelmasta on ladattu aineistoja lähes 620 000 kertaa.

Voidaan perustellusti sanoa, että Fenno-Ugrican aineistoja on käytetty runsaasti. Vuoden 2016 loppuun mennessä kokoelmasta oli tehty 891 000 latausta. Suosituimpia aineistoja ovat olleet kausijulkaisut, erityisesti komin- ja udmurtinkieliset sanomalehdet. Fenno-Ugrican käyttöä voi seurata osoitteessa: http://fennougrica.kansalliskirjasto.fi/simplestats/

Fenno-Ugrica on ollut yksi Kansalliskirjaston suosituimpia verkkokokoelmia latausmäärältään. Vertailun vuoksi voidaan todeta, että Kansalliskirjaston ylläpitämästä Fragmenta membranea -tietokannasta ladattiin aineistoja vuoden 2016 aikana noin 215 000 kertaa ja kotimaiseen kirjallisuuteen keskittyneestä Klassikkokirjastosta 82 145 kertaa. Kansalliskirjaston omista digitaalisista kokoelmista ainoastaan Doriassa oleva Pienpainetekokoelma ja Sanomalehtiarkisto ovat Fenno-Ugricaa käytetympiä digitaalisia kokoelmia.

3 Hankkeen toteuttaminen

Tässä yhteydessä tarkastellaan hankkeen toteuttamista lähinnä kielen dokumentoinnin, henkilöstön, suunniteltujen apuvälineiden, joukkoistamisen ja viestinnän näkökulmasta.

3.1 Henkilöstö

Sukukielten digitointiprojekti toteutettiin kaksivuotisena, kun apurahahakemuksessa hanke oli esitetty toteutettavaksi vuosina 2014–2016. Syynä tähän oli haettua pienempi apurahamyöntö, mikä ei mahdollistanut henkilöstön palkkaamista kolmeksi vuodeksi.

Sukukielten digitointiprojekti toteutettiin pääsääntöisesti projektihenkilöstön voimin. Hankkeeseen palkattiin projektipäällikkö, kaksi tietojärjestelmä asiantuntijaa ja kirjastosihteeri määräaikaisilla sopimuksilla. Koneen Säätiölle jätetyssä apurahahakemuksessa oli tavoitteena mainittu yhden tietojärjestelmäasiantuntijan ja kahden kirjastosihteerin palkkaaminen, mutta edellä mainitusta poikettiin lähinnä projektissa tarvittavien osaamisten vuoksi. Tekstieditorin kehitys ei pilottihankkeen aikana ollut edennyt silloin käytettävissä olleilla resursseilla toivotusti, joten projektiin päätettiin palkata toinen tietojärjestelmäasiantuntija. Tämä vaikutti keskeisellä tavalla projektin toteuttamiseen, henkilöstön palkkamiseen ja määräaikaisten työsuhteiden pituuteen.

Projektipäällikkö palkattiin ajalle tammikuu 2014-joulukuu 2015, kirjastosihteeri aloitti työnsä yksivuotisella sopimuksella keväällä 2014 ja sitä jatkettiin aina vuoden 2015 loppuun saakka palveluostoista säästyneillä varoilla. Kaksi tietojärjestelmäasiantuntijaa palkattiin hankkeeseen niin ikään vuoden sopimuksella vuoden 2014 alusta, mutta heidän sopimuksiaan pystyttiin jatkamaan palveluostoista säästyneiden varojen avulla aina elokuun 2015 loppuun saakka.

Palkkoja ja niiden sivukuluja maksettiin projektin aikana 281 702 euroa, mikä oli apurahahakemuksessa esiteltyä kuluarviolta (354 000 euroa) pienempi.

Projektipäällikkö sijoittui Kansalliskirjaston Tutkimuskirjastopalveluihin ja hän vastasi projektin toteuttamisesta sekä ohjasi hankkeen parissa työskentelevien henkilöiden työsuoritusta. Projektipäällikkö vastasi Venäjän Kansalliskirjastossa (Pietari)

tehtävän digitoinnin laatumääritysten noudattamisesta. Hän valvoi aineiston käyttöön asettamista Kansalliskirjaston julkaisujärjestelmässä ja aineistoa koskevien tekijänoikeussäädösten noudattamista sekä koordinoi Kansalliskirjastossa suoritettavia työvaiheita. Projektipäällikkö vastasi niin ikään sopimuksien valmistelusta ulkomaalaisten partnerien kanssa ja raportoi työstä sekä Kansalliskirjastolle että hankkeen ohjausryhmälle. Projektipäällikkö vastasi myös hankkeen viestinnän suunnittelusta ja osittain sen toteuttamisesta.

Projektissa työskennelleen kirjastosihteerin pääsääntöisinä työtehtävinä olivat digitoidun aineiston bibliografinen kuvailu ja sen käyttöön asettaminen Fenno-Ugrica -kokoelmassa. Lisäksi kirjastosihteeri vastasi venäjänkielisestä viestinnästä.

Projektissa työskennelleiden kahden tietojärjestelmäasiantuntijan työtehtävät liittyivät ennen kaikkea Revizor-oikolukujärjestelmään, sen suunnittelemiseen ja toteuttamiseen. Toinen tietojärjestelmäasiantuntijoista vastasi myös digitoidun aineiston jälkikäsittelystä ja Revizorissa käytettävien Alto XML-tiedostojen luomisesta, sillä Mikkelissä sijaitsevassa Kansalliskirjaston Digitointi- ja konservointikeskusessa ei voitu tätä työvaihetta suorittaa.

3.2 Revizor

Sukukielten digitointiprojektilla on kiinnekohtia myös kieliteknologiseen tutkimukseen, sillä projektin yhdeksi päämääräksi voidaan laajasti ajatella digitaalisten kirjasto- ja arkistoaineistojen käyttötapojen ja käytettävyyden parantaminen. Projektissa on edistetty suomalais-ugrilaisen aineiston käyttöön saattamisen lisäksi menetelmiä, joilla digitoitua raakadataa voidaan jalostaa entistä käyttökelpoisemmiksi aineistoiksi ja joilla aineistoa voidaan hyödyntää.

Sukukielten digitointiprojektissa näillä menetelmillä tarkoitetaan digitoidun aineiston OCR-tunnistuksen lisäämistä, tunnistetun tekstimassan palstoittamista sekä ennen kaikkea kielenkorjaukseen tarkoitetun OCR-editorin kehittämistä, jonka avulla voidaan digitoinnin ja OCR-tunnistuksen yhteydessä jääneitä virheitä korjata tehokkaasti ja talkoistamalla.

Jo Sukukielten digitointiprojektin pilottivaiheessa Kansalliskirjasto aloitti tekstinmuokkaamista helpottavan web-käyttöliittymän toteutuksen. Jatkohankkeen aikana tekstieditoria kehitettiin edelleen ja sen kehitystyö saatiin päätökseen kesällä 2015.

Tekstieditori Revizor koostuu kahdesta pääosasta, 1) editorin käyttöliittymästä, jota tekstin korjaajat käyttävät sekä 2) sen taustalla olevasta järjestelmästä, jossa hallinnoidaan tietokantoja, aineistoja ja niiden versiointia, käyttäjiä, selausnäkymiä ja muita editoinnin vaatimia toimintoja.

Käyttöliittymä on toteutettu JavaScriptillä ja taustajärjestelmä Pythonilla. Tiedonsiirto tapahtuu JSON API:lla toteutetun REST-rajapinnan avulla. Kansalliskirjasto julkaisi Revizorin avoimena lähdekoodina jatkohankkeen aikana. OCR-editoriin ladataan saaduista paketeista ALTO XML -tiedosto ja kuvatiedostot sekä paketeista saatava metadata; nimeke, tekijä ja pääasiallinen kieli. Näistä muodostetaan aineistoluettelo, josta aineisto voidaan valita avattavaksi editointikäyttöliittymässä. Ks.tekstieditori Revizorin dokumentaatio https://www.kiwi.fi/display/ocreditori

3.3 Joukkoistaminen

Aineistojen digitoimisen ja käyttöön saattamisen lisäksi hankekauden yksi päätavoite oli saada kielentutkijoiden käyttöön korjattuja kieliaineistoja. Sukukielten digitointiprojektissa talkoistaminen ja joukkoistaminen (crowdsourcing) pitäisi ymmärtää käsitteiden ”nichesourcing” ja ”kansalaistiede” kautta. Projektissa digitoidun aineiston luonteen ja siihen kohdistuvan tutkimuksen luonteen vuoksi joukkoistaminen suuren yleisön voimin ei tullut kysymykseen, sillä joukkoistettavien asiakirjojen määrä oli verrattain suuri. Projektissa valittu lähestymistapa korjata kieliaineistoja on pikemminkin ymmärrettävä nichesourcingina eli kohdennettuna tai tarvehakuisena joukkoistamisena kuin perinteisenä talkoistamisen muotona.

Nichesourcingin avulla pyrittiin yhdistämään ammattilaisten osaamista ja tiettyjen monimutkaistenkin työtehtävien ihmislähtöistä suorittamista. Tätä pyrittiin tavoittamaan valistuneiden, kieltä hallitsevien talkoistajien avulla. Samalla pyrittiin antamaan myös työn tekijälle näkyvät kasvot. Tämä ei ole aina ollut mahdollista perinteisissä talkoistamishankkeissa, joissa usein joukkoistamisen tavoitteena on ollut saada prosessoitua suuria määriä aineistoja, ehkä laadunkin kustannuksella.

Sukukielten digitointiprojektissa pyrittiinkin työskentelemään sellaisten henkilöiden ja yhteisöjen kanssa, joilla on jokin intressi aineiston suhteen ja joiden tavoitteena oli vaikkapa oman sukukielen hallinnan parantaminen. Sukukielten digitointiprojektin osalta tämä tarkoitti sitä, että pyrimme löytämään sellaisia partnereita, mahdollisesti äidinkielisiä, joille aineiston editointi voidaan antaa tehtäväksi. Sukukielten digitointiprojektissa näitä henkilöitä ja yhteisöjä voidaan hyvällä syyllä kutsua kansalaistieteilijöiksi, sillä heitä ohjattiin käyttämään OCR-editoria ja heidät perehdytettiin ymmärtämään myös projektin kielitieteellisiä tavoitteita.

Perinteisessä talkoistamisessa laajemmat tavoitteet on pirstottu mikrotehtäviksi, joiden suorittamiseen ei välttämättä tarvita aihepiirin erityisosaamista tai -taitoja. Nichesourcingissa pyritään hyödyntämään kansalaistieteilijöiden taitoja, esimerkiksi kielitaitoa ja paikallisten perinteiden tuntemusta, ja tuottamaan laadullisia tuloksia ja mahdollisesti uutta tietoa vaikkapa paikallisten murteiden ja traditioiden saralta. Tällaisen osaamisen hyödyntäminen kielitieteen hyväksi on tarpeen silloin,

kun ei ole välttämätöntä kartuttaa yleiskielen sanastoa, vaan tehtävänanto kohdennettiin jotain erityistä tarkoitusta varten.

Joukkoistamishankkeet eivät toteutuneet suunnitellussa laajuudessaan. Ainoastaan inkeroisen ja vepsänkielisiä teoksia pystyttiin oikolukemaan kieliyhteisöjen avustuksella, tällöinkin rahallista korvausta vastaan. Isoin ongelma Sukukielten digitointiprojektin kieliaineistojen oikoluvussa ja sen ohjaamisessa olivat sopivien yhteisöjen paikantaminen, työn merkityksen avaaminen oikolukijoille, ohjeistaminen ja motivointi. Sukukielten digitointiprojekteissa pyrittiin paikantamaan oikolukijoita eri kirjastojen ja korkeakoulujen kautta, mutta yhteydenotot eivät tuottaneet juurikaan tuloksia, mm. Oulussa lokakuussa 2014 järjestetyssä Suomalais-venäläisessä kulttuurifoorumissa.

Keväällä 2015, kun oli jo selvää, ettei projekti saa hankittua oikolukua laajemmissa määrin kielten puhujien keskuudesta, Sukukielten digitointiprojektissa päätettiin turvautua kotimaisten opiskelijoiden apuun ja Kansalliskirjasto tarjosi oikolukutyötä palkkiota vastaan. Oikolukukoulutukseen osallistui yhteensä 15 henkilöä, joista 12 työskenteli touko-kesäkuussa oikoluvun parissa. Tämän kampanjan seurauksena saimme oikoluettua hieman yli 10 000 sivua aineistoja eri kielillä. Oikoluetut aineistot julkaistiin uudelleen Fenno-Ugrica –kokoelmassa, samoin kuin oikoluetuista aineistoista luodut sanalistat.

Oikoluettu aineisto on myös linkitetty Kielipankin ylläpitämään Korp-konkordanssiohjelmaan, minkä avulla oikoluettua aineistoa voi käyttää kielentutkimuksen menetelmin.

4 Viestintä

Sukukielten digitointiprojektissa on pyritty hyödyntämään useita ulkoisia viestintäkanavia, sekä perinteisemmässä printtimuodossa että sosiaalisessa mediassa. Projektin viestintään on kuulunut keskeisenä osana myös tutkijayhteistyö, esitelmöiminen ja ohjausryhmätoiminta.

4.1 Viestintäkanavat verkossa

Sukukielten digitointiprojektissa on ollut käytössään Kansalliskirjaston sosiaalisen median viestintäkanavat ja projektilla on ollut omat suomen-, venäjän- ja englanninkieliset verkkosivut osoitteessa www.kansalliskirjasto.fi. Kansalliskirjaston verkkosivujen uudistuessa kevättalvella, tilaa venäjänkieliselle projektisivulle ei enää ollut, vaan projektilla oli enää suomen- ja englanninkieliset sivut.

Projekti on voinut viestiä Kansalliskirjaston Facebook-sivun kautta http://www.facebook.com/kansalliskirjasto , jossa on ilmoitettu mm. projektissa julkaistuista aineistoista ja projektiin liittyvistä tapahtumista. Projektipäällikkö on hyödyntänyt myös omaa henkilökohtaista twitter-tiliään viestiäkseen englanninkieliselle yleisölle keskeisistä projektin tapahtumista.

Projektin kannalta tärkeä viestintäkanava on ollut venäjänkielinen VKontakte-palvelu. Projektilla on ollut käytössään oma käyttäjätili http://vk.com/fennougrica, jonka avulla on voitu tavoittaa Venäjällä asuvia kieliyhteisöjä tehokkaasti. VKontakte-tili muodostui projektin aikana myös merkittäväksi kommunikaatiokanavaksi Venäjällä asuvien ja venäjää puhuvien tutkijoiden kanssa.

Projektilla on ollut oma englanninkielinen bloginsa http://blogs.helsinki.fi/fennougrica/, joka on toiminut keskeisimpänä projektiviestinnän kanavana. Projektiin ovat vuoroin kirjoittaneet niin tutkijat kuin projektissa työskentelevät henkilötkin.

Joitain Sukukielten digitointiprojektiin liittyviä kirjoituksia on julkaistu niin ikään Kansalliskirjaston Scripta Selecta -blogissa http://blogs.helsinki.fi/scriptaselecta/.

4.2 Esitelmät

Sukukielten digitointiprojektissa on pyritty jalkautumaan mahdollisimman lähelle tutkijaa ja tutkijoiden verkostoja. Sukukielten digitointiprojektin, siinä julkaistujen aineistojen ja verkostoitumisen kannalta on ollut ensiarvoisen tärkeää, että olemme

esitelleet työtämme sekä kirjastoalan, että fennougristiikan alan konferensseissa ja seminaareissa. Esitelmät löytyvät Kansalliskirjaston julkaistuarkistosta.

Sukukielten digitointiprojekti osallistui myös tieteellisen ohjelman järjestämiseen. 12. kansainvälinen fennougristikongressi järjestettiin Oulun yliopistossa 17.–21.elokuuta 2015 ja Sukukielten digitointiprojekti oli yhteisvastuussa yhden symposiumin järjestämisestä http://www.oulu.fi/suomenkieli/node/24466. Language Technology through Citizen Science -symposiumi käsitteli avoimen tieteen tilaa ja työkaluja uralilaisten kielten tutkimuksessa. Symposiumi oli yksi kongressin laajimmista.

Kansalliskirjasto osallistui myös Helsingin kirjamessuille syksyllä 2015. Kirjamessujen teemamaana oli tuolloin Venäjä ja Sukukielten digitointiprojektissa haluttiin tuoda messuilla esiin sekä Kansalliskirjastossa tehtyä työtä suomalais-ugrilaisten kielten tukemiseksi, että keskustella tiedeyhteistyöstä kielentutkimuksen parissa.

Kansalliskirjasto järjesti kirjamessuilla kaksi paneelikeskustelua, joista toinen keskittyi ylirajaiseen tiedeyhteistyöhön. Paneelissa oli mukana kielitieteen ja -teknologian, historiantutkimuksen ja kirjallisuudentutkimuksen ammattilaisia.

Lisäksi Kansalliskirjaston messuosastolla järjestettiin yhteensä 13 tutkijahaastattelua, joista viidessä keskusteltiin niin kieliteknologian merkityksestä uralilaisten kielten tutkimukselle kuin lähisukukielten asemastakin.

4.3 Printtimedia

Keskeisin printtimedia on ollut Kansalliskirjasto-lehti https://www.kansalliskirjasto.fi/fi/ajankohtaista/kansalliskirjasto-lehti, missä on toistuvasti julkaistu Sukukielten digitointiprojektiin liittyviä artikkeleja. Kansalliskirjasto-lehden lisäksi myös Scandinavian Library Quarterly julkaisi projektipäällikön artikkelin nichesourcingista http://slq.nu/?article=volume-47-no-4-2014-5

Painetun median merkitys projektin viestinnän kannalta voidaan katsoa olleen sähköistä viestintää vähäisempi.

4.4 Ohjausryhmä

Ohjausryhmätoiminta oli keskeinen ja tärkeä osa projektin tiedeviestintää. Ohjausryhmään kutsuttiin niin Kansalliskirjaston, rahoittajan kuin tiedeyhteisön edustajia, mikä mahdollisti sekä projektisuunnitelman valvonnan, että tiedonvaihdon kirjaston ja tutkijoiden kesken. Tämä keskusteluympäristö oli projektin kannalta merkittävä tiedonvälityksen kanava, jossa tutkijoita voitiin sitouttaa viestimään projektista eteenpäin, mutta myös saamaan tietoa tulevista fennougristiikan alan tapahtumista.

5 Apurahan käyttö

Koneen Säätiön myöntämä 650 000 euron apuraha jakaantui kustannuslajeittain seuraavasti:

			Yhteensä	Yhteensä	Yhteensä
			Budjetoidut kulut	Toteutuneet	Rah.jäljellä
WBS		Kustannuslaji	EUR	EUR	EUR
4703695	Koneen säätiö/Sukukielten digitointiproj	SAP yleinen_tiliryhm	650 043,69	20 325,50	-20 281,81
4703695	Koneen säätiö/Sukukielten digitointiproj	Tuotot		-650 000,00	
4703695	Koneen säätiö/Sukukielten digitointiproj	Kulut	650 043,69	670 325,50	-20 281,81
4703695	Koneen säätiö/Sukukielten digitointiproj	Palkat yhteensä	219 253,87	229 372,37	-10 118,50
4703695	Koneen säätiö/Sukukielten digitointiproj	Henkilösivukulut	49 654,42	52 330,47	-2 676,05
4703695	Koneen säätiö/Sukukielten digitointiproj	Tilakustannukset	0,00	3 052,00	-3 052,00
4703695	Koneen säätiö/Sukukielten digitointiproj	Aineet ja tarvikkeet	500,00	115,00	385,00
4703695	Koneen säätiö/Sukukielten digitointiproj	Koneet ja laitteet	1 373,06	1 544,30	-171,24
4703695	Koneen säätiö/Sukukielten digitointiproj	Poistot	0,00		0,00
4703695	Koneen säätiö/Sukukielten digitointiproj	Ostetut palvelut	267 053,44	266 420,76	632,68
4703695	Koneen säätiö/Sukukielten digitointiproj	Matkat	8 660,95	16 670,54	-8 009,59
4703695	Koneen säätiö/Sukukielten digitointiproj	Muut kulut	6 041,40	150,51	5 890,89
4703695	Koneen säätiö/Sukukielten digitointiproj	Yleiskustannukset	97 506,55	100 669,55	-3 163,00
Kokonaistulos			650 043,69	20 325,50	-20 281,81

Budjetti ylittyi 20 000 eurolla, johtuen lähinnä vuonna 2015 maksetuista palkoista, tilakustannuksista sekä matkoista.

6 Yhteenveto

Sukukielten digitointiprojekti oli Kansalliskirjastolle poikkeuksellinen hanke, sillä Kansalliskirjasto ei aikaisemmin ole profiloitunut erityisesti kielentutkimuksen tukiorganisaationa. Hyvät kontaktit venäläisiin partnereihin sekä hankkeen pilottivaiheessa (2012–2013), että jatkohankkeen aikana (2014–2016) mahdollistivat kansainvälisesti poikkeuksellisen projektin toteuttamisen.

Fenno-Ugricassa julkaistu aineisto on löytänyt hyvin yleisönsä niin tutkijoiden kuin kansalaistenkin keskuudessa. Fenno-Ugrica avattiin yleisölle keväällä 2013 ja siitä lähtien Sukukielten digitointiprojekteissa digitoituja aineistoja on ladattu 891 000 kertaa vuoden 2016 loppuun mennessä. Fenno-Ugrica on siis vakiinnuttanut asemansa uralilaisten kielten sähköisenä resurssina.

Aineistoja on käytetty myös tutkimuksen eri tarpeisiin useissa eri käyttöympäristöissä. Koska Fenno-Ugricassa julkaistun aineiston tekijänoikeudet ovat selvitetty ja aineistot on saatettu käytettäväksi kielen dokumentoinnin periaatteita noudattaen, on myös aineiston ja siitä luodun datan käyttö tutkijakunnalle luotettavaa.

Joukkoistaminen ja datan jalostaminen ei tuonut aivan toivottuja tuloksia. Onnistuneen joukkoistamisen esteenä oli etäisyys niihin kieliyhteisöihin, joita haluttiin osallistaa. Toisaalta myös maailmanpoliittinen tilanne sai useat venäläiset kansalaisjärjestöt ja korkeakoulut vetäytymään yhteistyöstä, etenkin Krimin valloituksen ja sitä seuranneiden lainsäädännöllisten muutosten vuoksi.

Koska joukkoistaminen ei tuottanut haluttuja tuloksia, niin Sukukielten digitointiprojektissa olisi pitänyt kääntää katse kielidatan jatkojalostamiseen toisin menetelmin. Tähän valitettavasti ei projektin henkilöstöllä ollut edellytyksiä, johtuen heidän osaamisensa painopisteistä, kun taas uuden, kieliteknologisempaa lähestymistapaa edustavan resurssin integroiminen hankkeeseen ei onnistunut.

Tutkijoiden integroituminen Sukukielten digitointiprojektiin avasi kirjastolle uusia yhteistyöverkostoja, joita sen on hyvä jatkossakin ylläpitää ja kehittää. Niin ikään Kansalliskirjaston luomat uudet kontaktit niin suomalais-ugrilaisille alueille kuin venäläiseen kirjastolaitokseen ovat hyvä voimavara, joihin tutkijoiden on hyvä jatkossakin nojautua.

Liitteet

Liite 1. Digitoidut aineistot.
Liite 2. Apurahahakemus Koneen Säätiölle.